NISTIR 4681

ON THE INTERCHANGEABILITY OF SGML AND ODA

Charles K. Nicholas
Lawrence A. Welsch

U.S. DEPARTMENT OF COMMERCE
National Institute of Standards
and Technology
Computer Systems Laboratory
Office Systems Engineering Group
Gaithersburg, MD 20899

U.S. DEPARTMENT OF COMMERCE
Rockwell A. Schnabel, Acting Secretary

NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
John W. Lyons, Director

NIST

NISTIR 4681

ON THE INTERCHANGEABILITY OF SGML AND ODA

Charles K. Nicholas
Lawrence A. Welsch

U.S. DEPARTMENT OF COMMERCE
National Institute of Standards
and Technology
Computer Systems Laboratory
Office Systems Engineering Group
Gaithersburg, MD 20899

January 1992

U.S. DEPARTMENT OF COMMERCE
Rockwell A. Schnabel, Acting Secretary

NATIONAL INSTITUTE OF STANDARDS
AND TECHNOLOGY
John W. Lyons, Director

On the Interchangeability of SGML and ODA

Charles K. Nicholas
Lawrence A. Welsch

Office Systems Engineering Group
Computer Systems Laboratory
National Institute of Standards and Technology
Gaithersburg, MD 20899

ABSTRACT

SGML and ODA are international standards for the markup and interchange of electronic documents. These standards are incompatible, in the sense that in general a document encoded using SGML cannot be used directly in an ODA-based system, and vice versa. We first describe these two standards, and suggest criteria under which a bridge between the two standards could be evaluated. We then evaluate the Office Document Language (ODL), an SGML application specifically designed for ODA documents, with respect to these criteria. We then describe a translation program that converts SGML documents to ODA and back.

Keywords: SGML, ODA, ODL, document interchange

Introduction

The Standard Generalized Markup Language (SGML)[1] and the Office Document Architecture (ODA) and Interchange Format[2] are the two most prominent standards for the markup and interchange of electronic documents. Considerable debate has ensued over the relative strengths and weaknesses of these two standards. Although they both address the problem of electronic document interchange, they differ in their approach to this problem, and in the functionality they provide to solve it.

The existence of two standards places those with interest in the document interchange problem in the unfortunate position of having to choose a standard, and thereby risk choosing the wrong standard, or accepting both (to some extent) and putting up with a duplication of user expertise and investment in interchange software. Communities of users have grown up around each standard, as well as companies committed to serving the software needs of those users. Any attempt by a government agency or standards organization to establish a single interchange standard by fiat will probably fail. Since it is reasonable to assume that both standards are going to remain in use for some time to come, it is appropriate to consider ways in which the duplication of effort costs can be minimized, and to find ways in which users of each standard can take advantage of the features offered by the other.

One way to reduce these costs and permit users to take advantage of both standards is to develop a mechanism whereby documents can be easily moved from one standard to the other. Furthermore, this mechanism should be as automatic as possible. Such a mechanism can then be regarded as a bridge between the two standards. If such a bridge were available, then

- Users of one standard would not need to learn the other in order to import or export documents encoded using the other standard. This would minimize duplication of effort in terms of user expertise.
- Developers of tools would be able to address the needs of a single market without significant additional investment.

- The purpose of both standards, document interchange, would be better served.

In this report, we propose some criteria that a bridge between SGML and ODA should satisfy, and assess a prospective bridge, known as the Office Document Language (ODL) with respect to these criteria. We then describe how SGML documents can be converted into ODA, and brought back again. We begin by discussing the capabilities of SGML and ODA at an intuitive level. (Brown[3] and Rosenberg, Sherman, Marks and Akkerhuis[4] present more detailed comparisons of the two standards.)

Two International Standards: SGML and ODA

SGML

The Standard Generalized Markup Language, or SGML, is a language and notation for describing classes of documents. In an SGML-encoded document, such as the document shown in Figure 1, the various elements of the document are delimited with distinguished character strings called *tags*. A document may, for example, have tags that delimit elements like paragraphs, subsections, appendices, figures, and so forth. A *start-tag* of the form <X> denotes the beginning of an element, and an *end-tag* of the form </X> denotes the end of that element.

```
<article>
<ti>On the Interchangeability of SGML and ODA</ti>
<au>Charles K. Nicholas
<au>Lawrence A. Welsch
<aff>Office Systems Engineering Group
 .
 .
 .
<sum>SGML and ODA are international standards for the markup and
 .
 .
 .
<sec>
<secti>Introduction</secti>
<para>The Standard Generalized Markup Language (SGML) <ref>1</ref> and
the Office Document Architecture (ODA) and Interchange Format
<ref>2</ref> are the two most prominent standards for the markup and
 .
 .
 .
</sec>
 .
 .
 .
</article>
```

Figure 1: An SGML-encoded document. The elements are delimited by tags, which appear inside angle brackets.

SGML provides several other types of markup, in addition to elements delimited by start-tags and end-tags. For example, a document may contain certain character strings, known as *entity references*,

which are replaced with designated values when encountered during the processing of the document.

In SGML, a class of documents is characterized by a grammar that indicates what markup is allowed, what markup is required, and how markup is distinguished from text. SGML defines this grammar with a *document type definition*, such as the one shown in Figure 2, that describes a class of documents in much the same way that formal grammars (written in BNF, for example) describe programming languages. In particular, document type definitions, or DTDs, describe how elements may be used in a document. The DTD in Figure 2 includes definitions for fifteen elements, indicating which start-tags and end-tags may be omitted and how the elements may be nested. The article element is delimited by the <article> start-tag and the </article> end-tag. The two dashes in the definition of the article element indicate that the start-tag and the end-tag must be present. The dash in the definition of para indicates that the start-tag must be present, and the "O" indicates that the end-tag may be omitted. The right side of an element definition contains information regarding the nesting of elements, in the form of a rule known as a *content model*, and may list exceptions to that rule, known as *inclusions* or *exclusions*. The content model for the article element, for example, indicates that the <article> start-tag must be followed by a title, one or more authors, an affiliation (presumably applying to all authors), a summary, and one or more sections. The symbol #PCDATA indicates a place where parsed character data (i.e. characters with no embedded markup) may appear.

SGML itself does not specify any particular set of elements; the set of elements that can be used in a document is specified by its document type definition. DTDs that describe complex, highly-structured documents, such as technical manuals for large systems, may have dozens of elements for the different components of the document.[5]

```
<!-- Sample Document Type Definition for Articles -->
<!ELEMENT article       - -     (ti, au+, aff, sum, sec+) >
<!ELEMENT sum           - -     (para+) >
<!ELEMENT sec           - -     (secti, (para+ | (para*, subsec+))) >
<!ELEMENT subsec        - -     (subsecti, para+) >
<!ELEMENT para          - O     (text | ref | fig)+ >
<!ELEMENT fig           - -     (graphic, figcap) >
<!ELEMENT (ti | au | aff | secti | subsecti | figcap | ref )
                        - O     (text) >
<!ELEMENT text          O O     (#PCDATA) >
<!ELEMENT graphic       - -     (#PCDATA) >
```

Figure 2: An SGML document type definition for simple articles. An SGML document type definition describes (among other things) the various tags to be used in marking up a document, the order in which they may appear, and whether they can be omitted.

An SGML-encoded document may be broken down into three parts: an *SGML declaration*, a *document type declaration*, and a *document instance*. The SGML declaration characterizes the document type declaration, and the subsequent document instance(s), in terms of character sets and optional features of SGML. The document type declaration may invoke a public document type definition, introduce an original document type definition (such as the one shown in Figure 2), or invoke a public DTD and add extra elements to it.

An SGML-encoded document is processed by parsing the document and performing appropriate actions, such as the generation of corresponding typesetter commands, when the various tags are recognized. There are two mechanisms in SGML for specifying the meaning of individual elements, in terms of processing to be performed, in the DTD:

- Application-specific *processing instructions* may be associated with the various elements in the DTD, and
- Markup substitution, i.e. the replacement of elements or entity references by other markup or text, can be specified using the SGML *link* feature.

Several organizations, such as the Association of American Publishers[6] and the U.S. Department of Defense[7] , have developed document type definitions with well-defined meanings for entity references and elements. (We have only touched on SGML's most important features. For a complete treatment of SGML, the reader is directed to Goldfarb's *SGML Handbook*.[8])

ODA

The Office Document Architecture and Interchange Format was designed to facilitate the interchange of documents in a heterogeneous network[9 ,10] and content of the document, as well as information on its appearance when rendered on a printed page (or other output media). For this reason, an ODA document is said to have a logical view and a layout view, as shown in Figure 3. These views correspond to trees, the leaves of which are associated with individual portions of the document's content.

The tree corresponding to the document's logical view is called the *specific logical structure*, and the root of this tree is the *document logical root*. The specific logical structure is shown in the top half of Figure 3. (The specific logical structure of this document is equivalent to the SGML document shown in Figure 1.) The *subordinates*, or descendants, of the document logical root may be of two types: composite logical objects and basic logical objects. The subordinates of composite logical objects may be other composite logical objects or basic logical objects. A basic logical object may be the parent of zero or more *content portions*. As an option, a logical object can be labeled with a sequence of ASCII characters known as the user-visible name attribute. Figure 3 shows the user visible names of the logical objects.

The layout view is represented, as shown in the lower half of Figure 3, in the form of a tree called the *specific layout structure*, rooted at the *document layout root*. The interior nodes of the specific layout structure represent specific portions of a document as it appears on an output media. Subordinates of the document layout root are layout objects. The types of ODA layout objects include *page set*, *composite page*, *basic page*, *frame* and *block*. Composite pages may contain (possibly nested) frames, and frames contain blocks. Individual content portions are associated with basic pages, or blocks within composite pages. The type of each layout object is shown in the lower half of Figure 3.

An ODA document belongs to one of three *document architecture classes*: the *formatted* document class, the *processable* document class, or the *formatted-processable* document class. A formatted form document is described in terms of the layout view. A processable form document is described in terms of the logical view. A formatted-processable form document, such as the document represented in Figure 3, is described in terms of both views.

The content of the document is shown between the specific logical and layout structures. Content portions may contain text; raster, facsimile or bit-mapped graphics information; or geometric graphics information in CGM format. Content portions may include special characters and other information that says, for example, how the content is to be rendered on an output device. Character, raster or bit-mapped graphics, and geometric graphics must conform to the ODA *content architectures* described in volumes 6, 7 and 8 of the ODA standard, respectively. (Additional volumes describing other content architectures are in various stages of preparation.)

In ODA, an *attribute* is a keyword and associated value that indicates some characteristic of that constituent or a relationship between that constituent and another constituent. There is a certain set of attributes associated with every constituent of an ODA document. For example, every specific logical object has an object identifier and an object type. Every composite object has a subordinates attribute, which lists the objects that appear below that composite object in the logical or layout view. The subordinates attributes of composite objects in the specific logical structure shows how a specific document is made up, in the same way that the nesting of tags describes the composition of SGML documents.

The various constituents of a document may have certain attribute values in common. These attributes may refer, for example, to specific formatting concerns such as indentation or point size. ODA provides two ways to specify the values of attributes for a number of objects in the same document:

- First, ODA has an elaborate mechanism for specifying default values of attributes. For example, the value of an attribute for a composite logical object may be used as the default value of that attribute for all composite and basic logical objects occurring below that object in the specific logical structure.
- Second, ODA allows a document to contain *generic logical objects* and *generic layout objects*. The attributes of generic objects need be specified only once, and may then be used repeatedly throughout a document. Once generic objects are defined and have values set for their attributes, a specific (logical or layout) object may declare itself to be a member of the class of objects defined by a given generic object, thereby inheriting the values of that generic object's attributes.

The set of generic logical objects in an ODA document makes up that document's *generic logical structure*, or GLS. The generic logical structure of an ODA document is similar to the document type definition for an SGML document, in that the generic logical structure shows how each part of the document is composed of smaller parts. Specifically, a composite object in the generic logical structure may have a *generator for subordinates* attribute that says, using a BNF-like notation, which types of object may appear as subordinates of that composite object. The generator for subordinates attribute is therefore similar to the content model of an SGML element, which lists the elements that may be nested inside that element. Generic logical structures are also similar to SGML DTDs in that a GLS can describe a class of documents, the members of which have the same structure but different content.

The appendix shows an example ODA document in the notation used by the ODA Toolkit developed for the EXPRES project.[4] The ODA document in the appendix belongs to the processable document class, so no layout view exists. Furthermore, no generic logical objects have been defined in this document. Each object is given a number of the form 1/nn as it is created. The content portions and basic objects were created before their parent objects so that the value of each object's subordinates attribute could be set immediately.

When a processable-form ODA document is to be rendered on a printed page or other output media, the objects in the specific layout structure need to be created. The ODA system performing this rendering may use *layout styles* associated with the document's specific or generic logical objects to construct the layout objects. The layout of the document's content portions can be specified through *presentation styles*. An organization may build customized generic logical and layout structures for its own use, with predefined content portions of various types and house layout and presentation styles.

Comparing SGML and ODA

The most important difference between SGML and ODA is their different stance towards the structure and processing of documents. SGML is a language for describing the structure of documents. The processing that occurs after a document is created is not addressed directly. In particular, the appearance of the document, before or after interchange, is not explicitly considered in SGML. At the moment, SGML incorporates no formatting information, although a complementary standard called DSSSL (Document Style Semantics and Specification Language) will be used to express formatting information. DSSSL is currently a Draft International Standard, with projected completion late in 1992.[11] Subsequent processing may be indicated with the SGML processing instruction syntax, which allows commands for some post-processor to be embedded in SGML markup.

ODA is less insistent than SGML about explicit description of document structure. ODA documents are self-describing, in two different senses:

1. As mentioned above, a given specific logical object has a subordinates attribute that lists the basic and composite logical objects which are that object's immediate descendants.
2. The subordinates of generic logical objects may be described using the generator for subordinates attribute.

The generic logical structure of an ODA document corresponds to the SGML notion of document type definition. ODA has a broader model of document processing than does SGML, a model that includes creation, editing, and formatting of a document as well as interchange. ODA incorporates formatting information directly into the document, through the specific layout structure and layout and presentation styles.

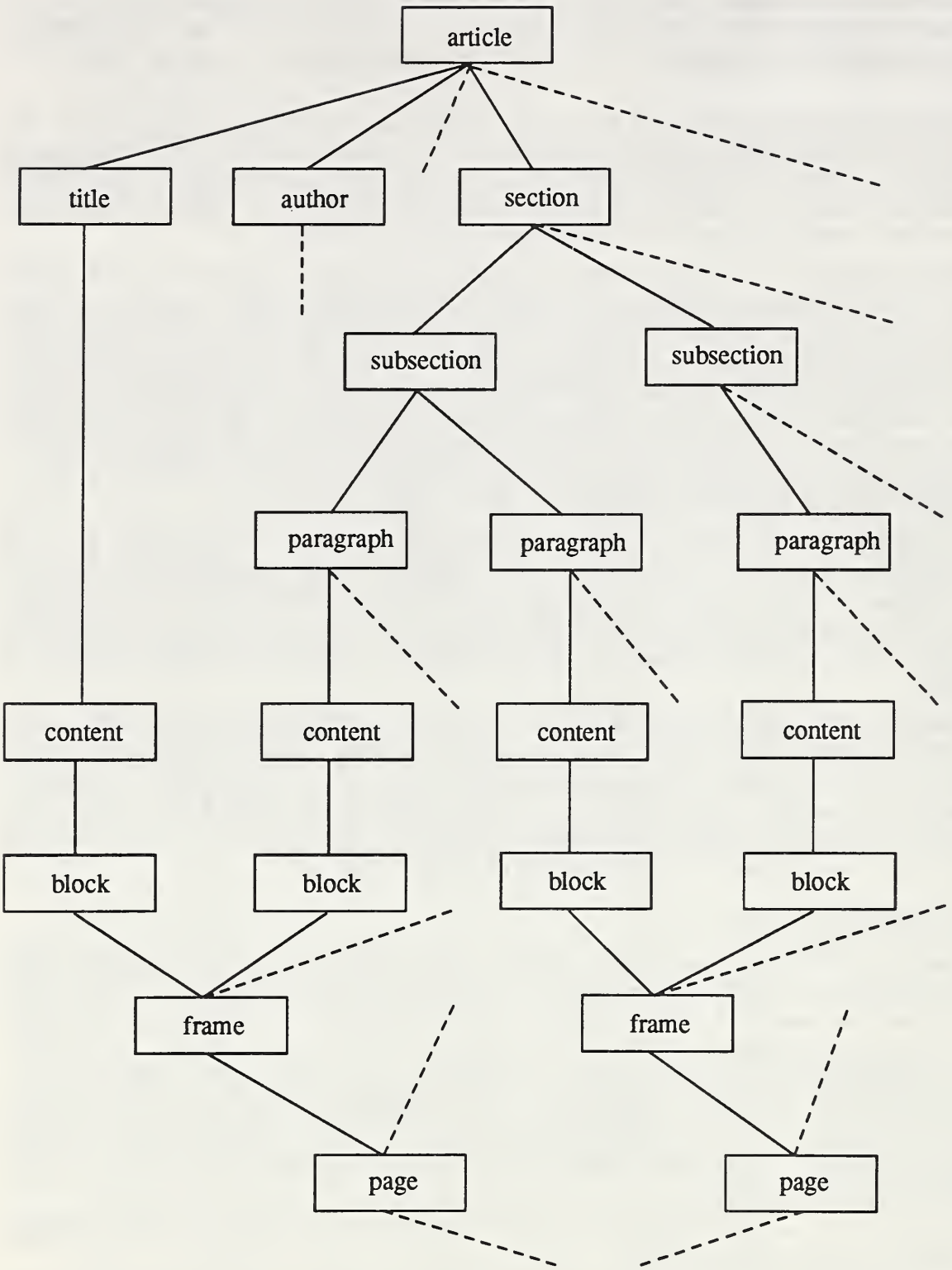

Figure 3: The Logical and Layout Views of a Document. An ODA document has a logical structure expressed in terms of, for example, sections, subsections, and paragraphs, as well as a layout structure, expressed in terms of blocks, frames, and pages.

SGML has gained acceptance in the publishing industry, where the ability to either define specialized document type definitions and formatting conventions, or use those prepared by others, suits individual user-companies that are accustomed to making their own decisions about a document's appearance when rendered. ODA was originally designed with a word processing environment in mind, assumes a more tightly-coupled community of users, and provides for organizational standards for content and structure as well as appearance. Their resulting differences in syntax and philosophy notwithstanding, both standards are capable of expressing the essential structure and content of a document.

Exchanging Documents Between the Standards

Communities of users have grown up around both SGML and ODA, along with tools and techniques peculiar to those communities. The standards themselves facilitate document interchange *within* these communities, but tools and techniques for the interchange of documents *between* these communities are not yet widely available or accepted. A tool that allows documents to be interchanged between SGML and ODA formats could then be regarded as a bridge, or gateway, between the two standards. Such a bridge should possess the following characteristics:

1. The bridge should be *reliable*, allowing for the migration of documents from one standard to the other with minimal (and ideally no) loss of information.
2. The bridge should be *automatic*, handling a wide variety of documents with minimal user intervention, but not at the expense of reliability.

The bridge may be used in two ways: for one-way translations of documents in which one standard is the source and the other is the target, and for two-way translations in which a document is moved from the source standard to the target with the intent of moving the document back to the source standard at some point in the future. A two-way translation can be viewed as the functional composition of two one-way translations. A reliable two-way translation must also preserve, or allow for the reconstruction, of any important information that might otherwise be lost as a result of either one-way translation.

Furthermore, a bridge between SGML and ODA needs to operate on a *structural* level as well as a *document* level. The SGML document type definition of a document contains structural information about a document (and, in fact, a class of documents) that should be preserved in the corresponding ODA document, whether as part of the specific logical structure of the ODA document, or as part of a generic logical structure that describes a set of ODA documents.

The content model of an element defined in an SGML DTD can most easily be captured in ODA as the value of the generator for subordinates attribute of a generic logical object. (Recall that the generator for subordinates attribute of an object specifies, in a BNF-like notation, the types of objects that can appear as subordinates of that object.) The generator for subordinates attribute is optional, according to the ODA standard. If the generator for subordinates attribute is defined for each of the generic logical objects that comprise the generic logical structure of an ODA document, then those generator for subordinates attributes are said to make up a *complete generator set.*

The essence of the translation problem at the structural level, then, is to determine whether the SGML DTD and the ODA generic logical structure, between which translation is to be done, are compatible. We say that an SGML DTD and an ODA GLS are *compatible* if and only if

1. A one-to-one correspondence exists between the elements in the SGML DTD and the object classes defined in the ODA GLS,
2. A complete generator set exists for the ODA generic logical structure, and
3. The content model for each SGML element is equivalent to the generator for subordinates attribute of the corresponding generic logical object.

Structural compatibility means that the content models in the SGML DTD and the complete generator set in the ODA generic logical structure are isomorphic, i.e. that the parse tree for an SGML document will be the same shape as the tree representation of the corresponding ODA document's logical view. Structural compatibility reduces the translation problem at the document level to a re-labeling of nodes in a tree, making it much easier to provide reliable automatic translation at the document level.

The converse of the above statement is also true: lack of structural compatibility makes it much harder to provide reliable automatic translation at the document level. If structural compatibility is lacking, however, all is not lost: A weaker notion of compatibility can be used for one-way translations on a document-by-document basis. A single ODA document is *sufficiently-compatible* with an SGML DTD if and only if

1. Each object class referenced in the ODA document corresponds to exactly one element in the SGML DTD, and
2. Each specific logical object appearing in the ODA document has a subordinates attribute (possibly derived from the generator for subordinates attribute of a generic logical object) that conforms to the content model of the corresponding SGML element.

Similarly, an SGML-encoded document and an ODA generic logical structure are sufficiently-compatible if and only if

1. Each element used in the SGML document corresponds to exactly one object class defined by an object in the ODA generic logical structure, and
2. The elements in the SGML document conform to the generator for subordinates attribute of the corresponding generic logical object.

We will see that if a document is sufficiently-compatible with a target structure, whether that target structure is an SGML DTD or an ODA GLS, translation of that particular document can still be done reliably and automatically. Figures 4 and 5 give an example ODA generic logical structure that is equivalent to the SGML DTD in Figure 2, using the notation defined in Part 2 of the ODA standard.

Several factors may interfere with structural compatibility:

- Structural compatibility is based on a one-to-one correspondence between the ODA document's logical object classes and the elements in the target SGML DTD. For example, such a correspondence exists between the elements in the SGML DTD shown in Figure 2 and the ODA generic logical structure shown in Figures 4 and 5. If no such correspondence exists, then significant user involvement may be needed to perform the conversions between tags and logical object classes.
- The SGML syntax for content models and the ODA syntax for the generator for subordinates attribute are similar, both being loosely based on BNF. They differ, however, in detail. For example, there is no easy way in ODA to specify inclusions or exclusions.
- SGML's *concurrent markup* feature allows a document to contain two or more sets of elements. The DTD could define two or more sets of elements to describe the document's logical structure. The DTD could also define a set of tags that describe the intended layout of the document in addition to a set of tags that describe its logical structure. An SGML document with two sets of elements would be structurally compatible with ODA if one set of SGML elements corresponded to the ODA logical view and the other set corresponded to the layout view. ODA does not currently support multiple logical views, making it difficult to achieve structural compatibility between ODA and an SGML document with two or more sets of logical elements. (It should be noted that concurrent markup is an optional feature of SGML, which not all SGML parsers support.)

So far, we have assumed the existence of a "target" structure, whether it be an SGML DTD or an ODA GLS, to which we are trying to convert. Unfortunately, when mapping an ODA document to SGML, a target DTD may not be available. If the ODA generator for subordinates attributes are present in the ODA document, the values of these attributes can be used to synthesize the content models of a target DTD. The generator for subordinates attribute, however, is an optional attribute of generic logical objects, so there is a very real chance that the document may not have a complete set of them. If an ODA document does not have all the generator for subordinates attributes (i.e. it does not have a complete generator set), we might still be able to synthesize a target DTD by using the generator for subordinates attributes that are available, if any, and by seeing which objects occur as subordinates of other objects in a given document's specific logical structure. For example, if we notice that objects of class A, B and C appear as subordinates of objects of class X, then the objects A, B and C can then be used to define for object class X a "catchall" generator for subordinates attribute of the form "OPT-REP(CHO(A, B, C)", which corresponds to the regular expression (and SGML content model) "(A | B | C)*".[12]

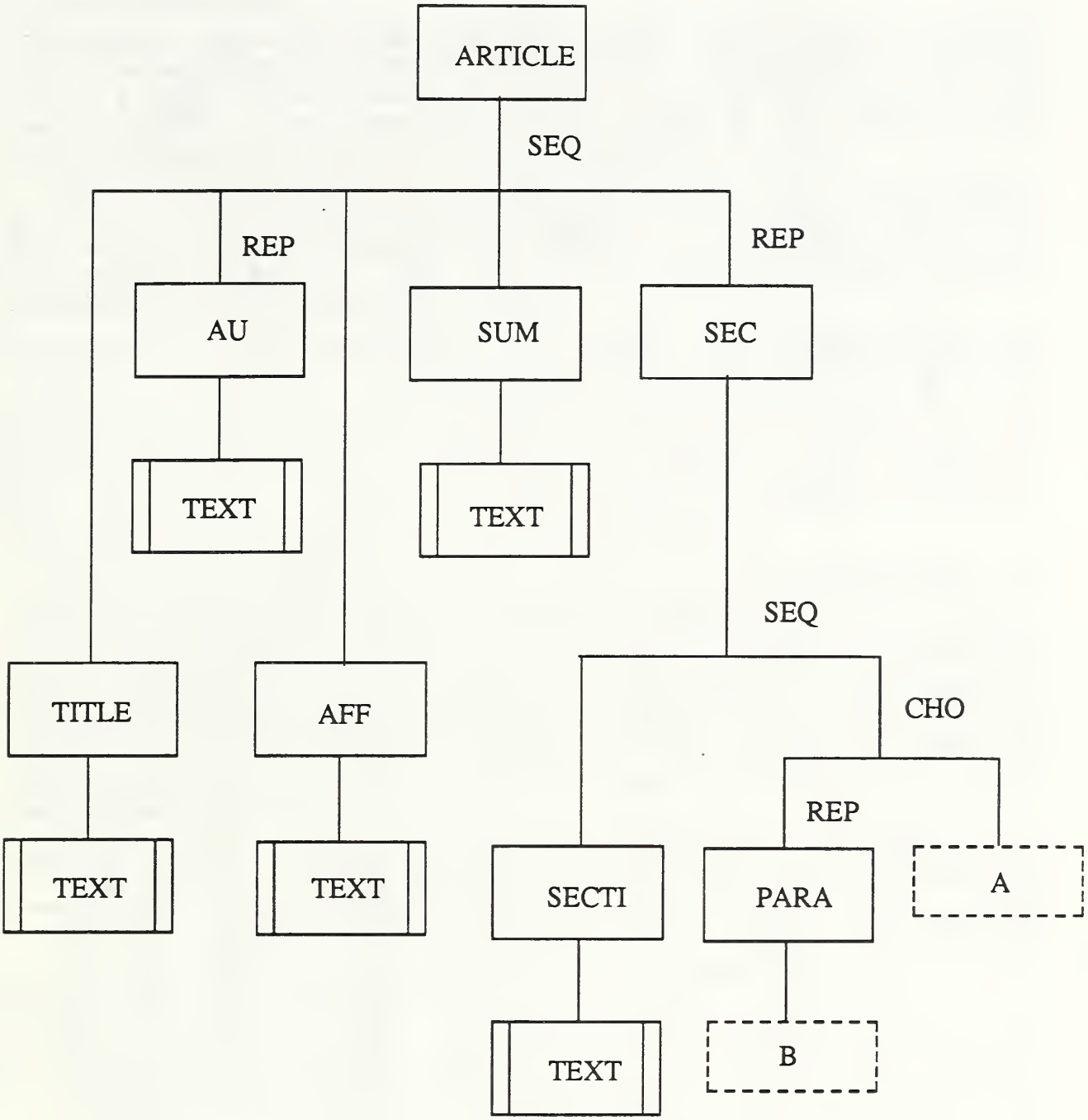

Figure 4: An ODA generic logical structure. This generic logical structure is structurally compatible with the SGML DTD in Figure 2.

As we mentioned above, if structural compatibility exists, then document level translation involves a mapping between SGML elements and ODA logical object classes. Even with structural compatibility, however, an ODA document in formatted or formatted-processable form carries layout information that is not easily transmitted in SGML without the use of the optional concurrent markup feature.

We mentioned earlier that a bridge between SGML and ODA should be both reliable and automatic. Given structural compatibility (including an appropriate target), automatic conversion at the document level is feasible. SGML documents can be parsed, and corresponding ODA objects generated, without human intervention. Similarly, it is not difficult to traverse the logical structure of an ODA document and generate

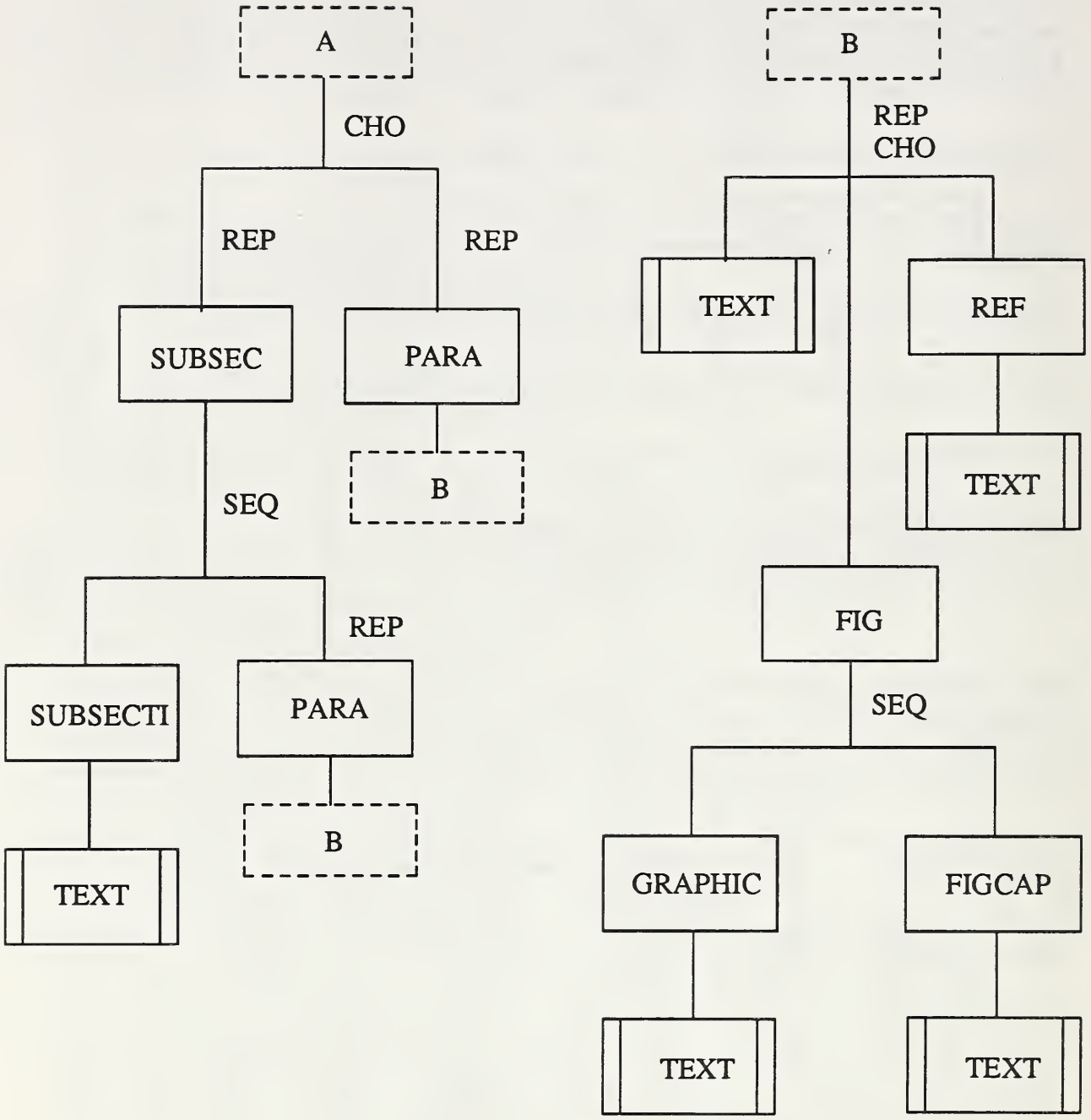

Figure 5: Sections consist of a sequence of subsections or a sequence of paragraphs. Paragraphs may contain text, figures, and bibliographic references.

corresponding SGML tags. The textual content in the ODA document will need to be scanned for special characters. For example, the character content architecture defines special characters to delimit subscripts and superscripts. Many SGML DTDs provide elements for the same purpose, and the mapping between these ODA special characters and the corresponding SGML tags would be straight-forward. Non-textual content, such as raster graphics, can be set off into separate files and brought into the SGML document as entity references or with the NDATA feature.

A Prospective Bridge: ODL

The ODA standard contains two formats that are meant to support document interchange. The first is the Office Document Interchange Format, or ODIF, which represents an ODA document as a binary encoding written in a general data structure description language known as ASN.1 (for Abstract Syntax Notation.)[13] The ASN.1 encoding is accompanied by an ASCII representation that is intended to be more easily read by humans. ASN.1 is described in detail in Tanenbaum.[14]

The other interchange format is ODL, or Office Document Language.[2] ODL is an SGML application (i.e. a set of SGML document type definitions) that allows ODA documents to be represented in an SGML format. ODL is equivalent to ODIF in the sense that both are capable of expressing the structure and content of an ODA document in detail. ODIF uses ASN.1 to express this information. ODL allows SGML elements to be used to describe the ODA document.

In ODL, the user-visible name attribute of an ODA object is used to generate the corresponding SGML tag. Consider, for example, the ODA document shown in the appendix. Each of the logical objects in that document, including basic and composite logical objects as well as the document logical root, has a user-visible name attribute corresponding to an element in the SGML DTD shown in Figure 2. As a result, when the ODA document in the appendix is converted to ODL, the output is the same as the SGML document shown in Figure 1. In case the (optional) user-visible name attribute has not been supplied, ODL provides a set of elements that describe the various constituents, such as basic logical objects, pages, frames, the document layout root, and so forth, that may exist in an ODA document.[2]

ODL allows ODA documents to be encoded in an SGML format. The ODL standard supplies DTDs for use with formatted, processable, and formatted-processable form documents. ODL was never intended to support translation of ODA documents to *arbitrary* SGML DTDs. When using ODL to translate an ODA source document to SGML, compatibility between the source document and the target DTD is the main concern. In particular, if the target DTD and the ODA document are not structurally compatible, the user becomes responsible for bringing the resulting ODL document into conformance with the target DTD. For example, if an object in the ODA document belongs to an object class with the user-visible name "affiliation", and the target DTD had no corresponding element, then human assistance would be required to complete the translation.

Two problems make it difficult to perform reliable translations from ODA to SGML using ODL:

1. When translating an ODA document (or class of documents) to SGML, the GLS of the ODA document is going to contain valuable structural information. In particular, it is not too hard to produce an SGML element definition that corresponds to any given ODA generic logical object. Unfortunately, the ODL standard currently says very little about how a generic logical structure is to be represented.
2. ODA objects in the specific logical structure need not be identified with particular object classes - and it is the class to which an object belongs that most closely corresponds to the SGML concept of element. An ODA document containing many such objects may still have an implicit structure in terms of sections, subsections, and so forth, but that structure is not explicit in the document's logical view, or in any attributes of the specific objects making up the document. When an ODA document lacking object class identifications or user visible names is mapped to ODL, the object type is used to create the required ODL tags: basic logical objects are labeled with ODL's <BLO> tag, composite logical objects are labeled with ODL's <CLO> tag, and the document logical root is labeled with the <DLOR> tag. Suppose, for example, that the user-visible name attributes had not been supplied for the objects in the ODA document in the appendix. Since that document has no generic logical structure from which default values of the user-visible name attribute could be derived, we would have to use the <DLOR>, <CLO>, or <BLO> tags in place of the <ARTICLE>, <SUM>, <PARA>, and other tags defined in Figure 2. This represents a significant loss of structural information.

A similar problem arises when translating from SGML to ODA: ODL makes no provision for encapsulating or otherwise preserving the structural information in the original DTD. No means of encapsulating the DTD of the original document is provided. Since ODL doesn't capture structural information, a translation system based solely on ODL must deal separately with the structural level issues of construction of a target DTD or GLS. This means that all sorts of structural information, including inclusions, exclusions, and concurrent markup, can be lost during translation from SGML to ODA via ODL. (Note that while

ODA does not support concurrent documents in general, ODL does. In fact, concurrent markup is used in the definition of ODL to capture the logical and layout views of a document simultaneously.) Because ODL gives very little insight into preserving the structural information contained in generic logical structures, and makes no provision at all for preserving the structural information in SGML DTDs, ODL currently does not satisfy the reliability property.

If the documents being translated are either structurally compatible or sufficiently-compatible, one-way translation from ODA to SGML via ODL can be done reliably and automatically. The ODL version of an ODA document can be produced by traversing the ODA document, writing out the textual content and the user visible names of object classes, in appropriate ODL syntax as they are encountered. If there is (at least) sufficient compatibility between the ODA document and the target DTD, each object class identifier can be mapped into a corresponding element in the target DTD. If no target DTD exists, one can be synthesized by using the information in the generic logical structure's generator for subordinates attributes, or the specific logical structure's subordinates attributes, as described above. Otherwise, some human assistance may be needed to map composite and basic logical objects to the correct elements in a target DTD.

One-way translation from SGML to ODA via ODL is slightly more complicated. If we assume sufficient compatibility between the SGML document and the target generic logical structure, then each SGML tag in the source document corresponds to an object class defined by a generic logical object in the target GLS. Then a given element name can be used verbatim in the corresponding ODL tag, which in turn can be used to compute the object class of the corresponding ODA object. Character data within tags can be mapped to character content inside a basic logical object. For example, if a <summary> element appears in the SGML source document, that tag can be written as is into the generated ODL document. When the ODL stream is converted to ODA, the object class defined by the generic logical object with the user-visible name "summary" can be identified, and that object class identifier can be attached to an ODA basic logical object containing the text of the summary.

In summary, the main obstacles to reliable translation between ODA and SGML via ODL are the loss of structural information in the DTD or GLS, and the loss of layout information in formatted and formatted-processable form ODA documents. Given that one-way translation between a compatible source and target is relatively easy to automate, two-way translations, which can be thought of as successive one-way translations, are also easy to automate.

Recommendations

Converting from SGML to ODA

We now consider the problem of designing a translation program that converts SGML documents to ODA with the intent of perhaps converting those documents back to SGML at some point in the future. If software to create ODA documents is available (the ODA Toolkit built as part of the EXPRES project provides this functionality[4]), then the translation routine may build the desired ODA document directly. Otherwise, the output of the translation program would be an ODL or ODIF data stream.

In two-way translation, the structural information in the source DTD must be preserved and transmitted with the document when it is sent to ODA, so that sufficient compatibility between the new ODA document and the original DTD can be maintained. If the translation program is designed to create the ODA or ODIF documents directly, without using ODL, then the source DTD can be preserved by converting the content models of the SGML elements into generator for subordinates attributes of corresponding ODA generic logical objects, thereby creating a target generic logical structure. If the translation program uses ODL, then the source DTD should be stored separately.

Once the target GLS is ready, the SGML-encoded document must be parsed. Three factors should be considered during the parsing process:

- The parser will need to intercept and resolve entity references involving local files. Entity references that resolve to strings of characters may appear in ODL documents, and should be placed in the character content portions of basic logical objects during the conversion from ODL to ODA. However, entity references may also refer to local files that contain non-character data, such as raster graphics or line drawings. The parser would need to record the data in these files, so that it can be copied into

the content portions of corresponding basic logical objects.

- Attributes may occur within SGML tags. The DTD describes the attributes that may be associated with specific elements. For example, a figure reference element may have the form <FIGREF ID=n>, where the value of the ID attribute is the number of the figure being referenced. SGML attributes can be simulated using the ODA bindings mechanism, which allows a name/value pair to be associated with a specific or generic object. When an attribute declaration appears in the definition of an SGML element, a binding can be defined for the corresponding generic logical object. A default value can be associated with the binding at the time of definition, and instances of that attribute value can be associated with corresponding specific logical objects as the attribute is encountered in SGML tags.
- No construct in ODA corresponds to the notion of marked sections in SGML. According to Goldfarb, "A marked section of a document is one that has been identified for a special purpose, such as ignoring markup within it."[8] One appropriate use of the marked section feature, in this context of document interchange, would be to select portions of the SGML document that are not to be converted to ODA.

Given that these caveats are observed, then the translation routine can parse the SGML document, copying content at the leaves of the parse tree into ODA content objects, and mapping elements into basic and composite logical objects.

Converting from ODA to SGML

When designing a program to convert a document from ODA to SGML, whether ODL is used or not, the main problem is to guarantee structural compatibility. Recall that ODA documents can be in formatted, processable, or formatted-processable form. Since the lack of structural information in a formatted-form document makes structural compatibility hard to attain, we restrict ourselves to documents in processable or formatted-processable form. We first consider the translation of processable form documents, and then discuss the extra steps needed to prevent loss of layout information during the translation of formatted-processable form documents.

If structural information is lacking, structural compatibility becomes harder to guarantee. There are four cases to consider:

1. If the ODA document's GLS and a target DTD are both available, we can test their structural compatibility by building a one-to-one correspondence between ODA object classes and SGML tags. If the GLS and the target DTD enjoy structural compatibility, then we can produce an ODL version of the ODA document, generating tags in the ODL document that conform to the target DTD.
2. If the GLS is available but not the target DTD, the generator for subordinates attributes in the generic logical structure can be used, as described above, to synthesize an appropriate target DTD.
3. If the target DTD is available but not the GLS, then human assistance will be needed to figure out which tags in the target DTD correspond to the various basic and composite logical objects in the ODA document.
4. If neither the target DTD or the GLS are available, then we can generate an ODL file using the <DLOR>, <CLO>, and <BLO> tags.

If no generic logical structure is available, then translation from ODA to SGML is more difficult. If we were willing to settle for partial automation, it would be relatively easy to build a general tool for mapping ODL tags to some other SGML DTD that would occasionally need to ask a human user which of several candidate tags (in the target DTD) is appropriate. If the target DTD had a regular structure, e.g. of the form title, author, sequence of paragraphs, then it would be easy to implement a translation routine as a finite state machine, for example. However, such a routine would not respond well (e.g. by generating tags that don't conform to the target DTD) to ODL documents that deviate from the pattern. A somewhat more specialized pattern-matching program (based, for example, on attribute grammars rather than finite state machines) would require more effort to build, but would be able to process conforming documents with relatively little human intervention. (This was the approach used in the Chameleon document translation system.)[15]

When ODL is used to encode a formatted-processable document, containing both logical and layout information, each content portion needs to be tagged in accordance with what are, in effect, two different SGML DTDs: One for describing logical structure, and the other for describing layout structure. If the concurrent markup feature is not available, one alternative approach is to use a single document encoded with logical tags, and generate a version of the same document with layout tags via the SGML link mechanism. The link mechanism performs naive tag substitution, however, and cannot make decisions based, for example, on the nesting level at which a given tag resides. The link mechanism requires *a priori* knowledge of how each tag is to be laid out, without regard to nesting level, or the values of SGML or ODA attributes.[5]

The Design and Implementation of ODL-Based Translators

We have written software that demonstrates the interoperability of ODA and SGML, using the ideas presented in this report. SGML_to_ODA converts SGML documents into ODA. ODA_to_SGML reads an ODA document from an ODIF file and converts that document to ODL.

SGML_to_ODA begins by reading a table listing the elements in the SGML document and the object class identifiers to be used for the corresponding ODA generic logical objects. This is used to create an ODA document that contains a structurally compatible generic logical structure, but which is otherwise empty. SGML_to_ODA then reads and parses an SGML-encoded document. During the parse, the content and SGML tags in the SGML document are converted to ODA objects in the specific logical structure. Each basic and composite logical object is assigned an object class identifier corresponding to a basic or composite logical object in the generic structure. SGML_to_ODA then creates an ODIF version of the ODA document, including the generic and specific logical structures, that can be saved or transmitted as needed. (The current version of SGML_to_ODA allows entity references to be present in the SGML document, but makes no attempt to resolve them.)

SGML_to_ODA was written in YACC, Lex and C. YACC and Lex were used to generate a small, DTD-specific SGML parser. The action routines in the parser make use of library calls in the ODA Toolkit[4] to build the various components of the ODA document. We could have also used a commercial or public-domain SGML parser to process the SGML document and make the necessary calls to the ODA Toolkit. We decided not to use the NBS parser described in [16] since that parser does not provide for action routines.

ODA_to_SGML begins by reading an ODIF file from disk. ODA Toolkit calls are made to convert the ODIF file into an internal ODA data structure. The specific logical structure of this ODA document is then traversed, left to right, in preorder. As the ODA document is traversed, the character content and user-visible names of the object classes appearing in the specific logical structure are written as an ODL output file.

These two programs were tested on an SGML version of this report. The SGML input file was 45701 bytes in length. Twenty-nine different SGML element definitions were used in the document, and several elements were set up to allow start-tags or end-tags to be omitted. Otherwise, no effort was made to minimize markup. When reading the SGML version of this report and writing the corresponding ODIF document, SGML_to_ODA ran at approximately 70 lines per second on a lightly-loaded SUN 4/280. The resulting ODIF file was 100624 bytes in length. The size increase can be explained by noting that the ODIF file contained descriptions of twenty-nine generic logical objects, and a number of ODA attributes, such as object identifier, that are implicit in the SGML document.

The ODA document was then converted back to SGML. The differences between the resulting SGML document and the original were due to SGML_to_ODA's treatment of whitespace. During the parse, some newline characters that occurred between tags in the original document, in locations where text was permitted, were interpreted as text and were placed in the ODA document as character content. Whitespace that occurred between tags in places where text was not allowed was ignored.

Conclusion

In this report, we have described the SGML and ODA standards, and proposed a set of criteria for evaluating a bridge mechanism between SGML and ODA. We saw that if the documents being exchanged are structurally or sufficiently compatible with the target standard, then the criteria of reliability and

automaticity can be satisfied. ODL's weakness as a bridge comes from its lack of a mechanism to preserve the structural information in an SGML DTD or an ODA generic logical structure.

When planning to convert documents from one standard to the other, it is important to make sure that the source document(s) and the target standard are compatible. SGML features that can interfere with structural compatibility, such as inclusions, exclusions, and concurrent markup, should be used with care. ODA documents should be represented in processable or formatted-processable form, so that explicit structural information in the form of generic logical objects or user-visible names can be preserved. If ODL is to be used for interchange, then the structural information in the source document, in the form of an SGML document type definition or an ODA generic logical structure, should be maintained.

We have shown that a translator based on the idea of structural compatibility can operate reliably with minimal user intervention. In addition, our experience suggests that under some circumstances SGML may be preferable to ODIF as an interchange format for ODA documents, due to the compactness of SGML files in comparison to ODIF, and the fact that SGML files can be read by humans and ODIF files cannot.

We expect ODA and SGML to coexist in the typical office environment. ODA's ability to handle logical and layout components of a document should make ODA compatibility an important feature of document processing systems in the future. However, the tasks of authoring and formatting/composing are frequently separated in practice. This is the case, for example, in traditional publishing. At the moment, the larger number of SGML-compatible systems will cause SGML to remain the standard of choice in such situations, but this will change as the number of ODA-compatible systems increases.

Acknowledgements

We extend our thanks to John Barkley, Frank Dawson, Fran Nielsen, Sandy Mamrak, Judi Moline, Mark Sherman, and the anonymous referees for their comments on earlier versions of this report.

References

1. *Information Processing - Text and Office Systems - Standard Generalized Markup Language (SGML)*, October 1986. Ref. No. ISO 8879-1986(E). First edition - 1986-10-15. FIPS PUB 152.
2. *Information Processing - Text and Office Systems - Office Document Architecture (ODA) and Interchange Format*, 1989. ISO 8613 Parts 1-8.
3. Heather Brown, "Standards for Structured Documents," *The Computer Journal*, vol. 32, no. 6, pp. 505-515, December 1989.
4. Jonathan Rosenberg, Mark Sherman, Ann Marks, and Jaap Akkerhuis, *Multi-media Document Translation: ODA and the EXPRES Project*, Springer-Verlag, New York, 1991.
5. Martin Bryan, *SGML: an author's guide to the Standard Generalized Markup Language*, Addison-Wesley, 1988.
6. Association of American Publishers, *Reference Manual on Electronic Manuscript Preparation and Markup*, EPSIG c/o OCLC, Dublin, OH 43017-0702, 1989.
7. *Technical Manuals: Markup Requirements and Generic Style Specification for Electronic Printed Output and Exchange of Text*, U.S. Department of Defense, July 1990.
8. Charles F. Goldfarb, *The SGML Handbook*, Oxford University Press, 1991.
9. Wolfgang Horak, "Office Document Architecture and Office Document Interchange Formats: Current Status of International Standardization," *IEEE Computer*, vol. 18, no. 10, pp. 50-60, October, 1985.
10. R. Hunter, P. Kaijser, and F. Nielsen, "ODA: a document architecture for open systems," *Computer Communications*, vol. 12, no. 2, pp. 69-79, April 1989.
11. *Final Text, ISO/IEC CD 10179, Information Technology - Text and Office Systems - Document Style Semantics and Specification Language (DSSSL)*, 1991.
12. Golkar, Seyed N., *Experience with Document Interchange*, IBM/NSF Workshop on Compound Document Exchange using ODA, 1989.
13. *Information Processing - Open Systems Interconnection - Specification of Abstract Syntax Notation One (ASN.1)*, 1987. ISO 8824
14. Tanenbaum, Andrew S., *Computer Networks*, Prentice-Hall, Englewood Cliffs, NJ, 1988.
15. Mamrak, S., Kaelbling, M., Nicholas, C., and Share, M., "Chameleon: A System for Solving the Data Translation Problem," *IEEE Transactions on Software Engineering*, vol. 15, no. 9, pp. 408-414, September 1989.
16. Jim Heath and Larry Welsch , "Difficulties in Parsing SGML," in *Proceedings of the ACM Conference on Document Processing Systems*, pp. 71-77, ACM Press, Santa Fe, New Mexico, December, 1988.

Appendix

The following listing shows a portion of an ODA document that is equivalent to the SGML document shown in Figure 1. Each constituent of the ODA document is shown with its required attributes. The user-visible name attribute is optional.

```
Specific Logical Structure:

1/1: SPECIFIC DOCUMENT LOGICAL ROOT
     ("Object Identifier" "3")
     ("Object Type" 'document logical root')
     ("Subordinates" 1/3: 1/5: 1/7: 1/9: 1/12: 1/15: ... )
     ("User-visible Name" "article")

1/3: SPECIFIC BASIC LOGICAL OBJECT
     ("Content Architecture Class" 2 8 2 6 1
         ['processable character content architecture'])
     ("Content Portions" 1/2:)
     ("Object Identifier" "3 0")
     ("Object Type" 'basic logical object')
     ("User-visible Name" "ti")

1/2: CONTENT PORTION (character)
     ("Content Identifier - Logical" "3 0 0")
     ("Content Information" 'On the Interchangeability of SGML and ODA')
     ("Type of Coding" 2 8 3 6 0
         ['character content encoding'])

1/5: SPECIFIC BASIC LOGICAL OBJECT
     ("Content Architecture Class" 2 8 2 6 1
         ['processable character content architecture'])
     ("Content Portions" 1/4:)
     ("Object Identifier" "3 1")
     ("Object Type" 'basic logical object')
     ("User-visible Name" "au")

1/4: CONTENT PORTION (character)
     ("Content Identifier - Logical" "3 1 0")
     ("Content Information" 'Charles K. Nicholas')
     ("Type of Coding" 2 8 3 6 0
         ['character content encoding'])

1/7: SPECIFIC BASIC LOGICAL OBJECT
     ("Content Architecture Class" 2 8 2 6 1
         ['processable character content architecture'])
     ("Content Portions" 1/6:)
     ("Object Identifier" "3 2")
     ("Object Type" 'basic logical object')
     ("User-visible Name" "au")
```

```
1/6: CONTENT PORTION (character)
    ("Content Identifier - Logical" "3 2 0")
    ("Content Information" 'Lawrence A. Welsch')
    ("Type of Coding" 2 8 3 6 0
        ['character content encoding'])

1/9: SPECIFIC BASIC LOGICAL OBJECT
    ("Content Architecture Class" 2 8 2 6 1
        ['processable character content architecture'])
    ("Content Portions" 1/8:)
    ("Object Identifier" "3 3")
    ("Object Type" 'basic logical object')
    ("User-visible Name" "aff")

1/8: CONTENT PORTION (character)
    ("Content Identifier - Logical" "3 3 0")
    ("Content Information" 'Office Systems Engineering ...')
    ("Type of Coding" 2 8 3 6 0
        ['character content encoding'])

1/12: SPECIFIC COMPOSITE LOGICAL OBJECT
    ("Object Identifier" "3 4")
    ("Object Type" 'composite logical object')
    ("Subordinates" 1/11:)
    ("User-visible Name" "sum")

1/11: SPECIFIC BASIC LOGICAL OBJECT
    ("Content Architecture Class" 2 8 2 6 1
        ['processable character content architecture'])
    ("Content Portions" 1/10:)
    ("Object Identifier" "3 4 0")
    ("Object Type" 'basic logical object')
    ("User-visible Name" "para")

1/10: CONTENT PORTION (character)
    ("Content Identifier - Logical" "3 4 0 0")
    ("Content Information" 'SGML and ODA are international standards ...')
    ("Type of Coding" 2 8 3 6 0
        ['character content encoding'])

1/15: SPECIFIC COMPOSITE LOGICAL OBJECT
    ("Object Identifier" "3 5")
    ("Object Type" 'composite logical object')
    ("Subordinates" 1/14:)
    ("User-visible Name" "sec")
```

```
1/14: SPECIFIC BASIC LOGICAL OBJECT
    ("Content Architecture Class" 2 8 2 6 1
        ['processable character content architecture'])
    ("Content Portions" 1/13:)
    ("Object Identifier" "3 5 0")
    ("Object Type" 'basic logical object')
    ("User-visible Name" "secti")

1/13: CONTENT PORTION (character)
    ("Content Identifier - Logical" "3 5 0 0")
    ("Content Information" 'Introduction ')
    ("Type of Coding" 2 8 3 6 0
        ['character content encoding'])

 .
 .
 .
```

NIST-114A
(REV. 3-90)

U.S. DEPARTMENT OF COMMERCE
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY

BIBLIOGRAPHIC DATA SHEET

1. PUBLICATION OR REPORT NUMBER
NISTIR 4681

2. PERFORMING ORGANIZATION REPORT NUMBER

3. PUBLICATION DATE
JANUARY 1992

4. TITLE AND SUBTITLE

On the Interchangeability of SGML and ODA

5. AUTHOR(S)

Charles K. Nicholas and Lawrence A. Welsch

6. PERFORMING ORGANIZATION (IF JOINT OR OTHER THAN NIST, SEE INSTRUCTIONS)

U.S. DEPARTMENT OF COMMERCE
NATIONAL INSTITUTE OF STANDARDS AND TECHNOLOGY
GAITHERSBURG, MD 20899

7. CONTRACT/GRANT NUMBER

8. TYPE OF REPORT AND PERIOD COVERED

9. SPONSORING ORGANIZATION NAME AND COMPLETE ADDRESS (STREET, CITY, STATE, ZIP)

NIST/CSL
225/B266
Gaithersburg, Maryland 20899

10. SUPPLEMENTARY NOTES

11. ABSTRACT (A 200-WORD OR LESS FACTUAL SUMMARY OF MOST SIGNIFICANT INFORMATION. IF DOCUMENT INCLUDES A SIGNIFICANT BIBLIOGRAPHY OR LITERATURE SURVEY, MENTION IT HERE.)

SGML and ODA/ODIF are international standards for the markup and interchange of electronic documents. These standards are incompatible, in the sense that a document encoded using SGML cannot be used directly in an ODA-based system, and vice versa. We first describe these two standards, and suggest criteria under which a bridge between the two standards could be evaluated. We then evaluate the Office Document Language (ODL), an SGML application specifically designed for ODA/ODIF documents, with respect to these criteria. We then describe a translation program that converts SGML documents to ODA and back.

12. KEY WORDS (6 TO 12 ENTRIES; ALPHABETICAL ORDER; CAPITALIZE ONLY PROPER NAMES; AND SEPARATE KEY WORDS BY SEMICOLONS)

conversion; electronic documents; ODA/ODIF; ODL; SGML; translation

13. AVAILABILITY

XXX	UNLIMITED
	FOR OFFICIAL DISTRIBUTION. DO NOT RELEASE TO NATIONAL TECHNICAL INFORMATION SERVICE (NTIS).
	ORDER FROM SUPERINTENDENT OF DOCUMENTS, U.S. GOVERNMENT PRINTING OFFICE, WASHINGTON, DC 20402.
X	ORDER FROM NATIONAL TECHNICAL INFORMATION SERVICE (NTIS), SPRINGFIELD, VA 22161.

14. NUMBER OF PRINTED PAGES
22

15. PRICE
A02

ELECTRONIC FORM